JOURNAL OF CHEMOMETRICS
J. Chemometrics **13,** 295–309 (1999)

PARAFAC2—PART II. MODELING CHROMATOGRAPHIC DATA WITH RETENTION TIME SHIFTS

RASMUS BRO,[1]* CLAUS A. ANDERSSON[1] AND HENK A. L. KIERS[2]

[1]*Chemometrics Group, Food Technology, Department of Dairy and Food Science, The Royal Veterinary and Agricultural University, Denmark*

[2]*Heymans Institute (PA), University of Groningen, Groningen, The Netherlands*

SUMMARY

This paper offers an approach for handling retention time shifts in resolving chromatographic data using the PARAFAC2 model. In Part I of this series an algorithm for PARAFAC2 was developed and extended to N-way arrays. It was discussed that the PARAFAC2 model has a number of attractive features. It is unique under mild conditions though it puts fewer restrictions on the data than the well-known PARAFAC1 model. This has important implications for the modeling of chromatographic data in which retention time shifts can be regarded as a violation of the assumption of parallel proportional profiles underlying the PARAFAC1 model. The PARAFAC2 model does not assume parallel proportional elution profiles, but only that the matrix of elution profiles preserve its 'inner-product structure' from sample to sample. This means that the cross-products of the matrix holding the elution profiles in its columns remain constant. Here an application using chromatographic separation based on the molecular size of thick juice samples from the beet sugar industry illustrates the benefit of using the PARAFAC2 model. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS: multiway; curve resolution; fluorescence spectroscopy; shifted profiles

INTRODUCTION

In order to understand the chemistry of the color formation during sugar processing from beets, an experiment was conducted to explore the presence and amount of chemical analytes in thick juice, which is an intermediate product in the sugar production. The molecular entities of thick juice samples were separated by size and affinity on a chromatographic system and detected by fluorescence in the hope that the individual fluorophores could be separated and detected. However, it turned out to be impossible to separate the analytes completely; that is, the elution peaks/profiles were partly overlapping. The analysis was further complicated by the fact that there were huge shifts in retention time of specific analytes from sample to sample.

Overlapping chromatographic peaks can sometimes be separated mathematically. If a univariate detection system is used in a chromatographic system, an experiment results in a time profile which is conveniently held in a vector. If several such experiments are performed on different samples, a matrix **X** results, of which each row holds the profile of each individual sample. If there are no

* Correspondence to: R. Bro, Chemometrics Group, Food Technology, Department of Dairy and Food Science, The Royal Veterinary and Agricultural University, Denmark. E-mail: rb@kul.dk

Contract/grant sponsor: Nordic Industry Foundation; Contract/grant number: P93149
Contract/grant sponsor: FØTEK

CCC 0886–9383/99/030295–15 $17.50

Received 25 July 1998
Revised 6 January 1999
Accepted 18 February 1999

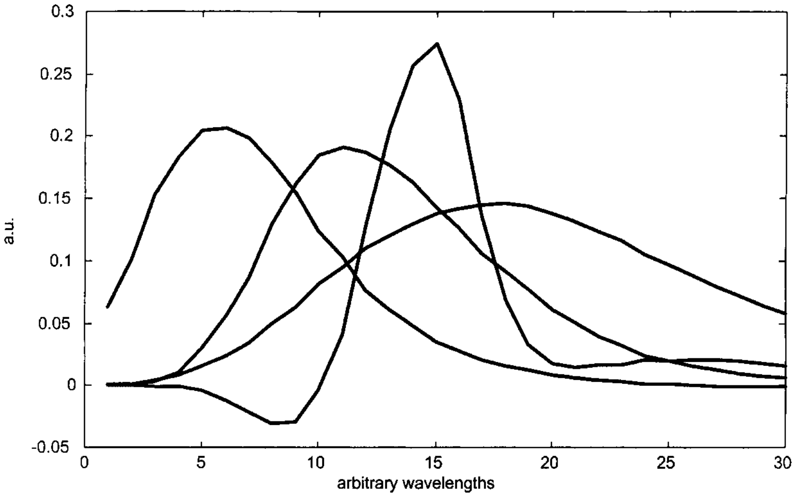

Figure 1. Spectra used in simulated data

retention time shifts in the data, every analyte will give rise to the same elution profile in every sample, except for a change in magnitude (area) depending on the concentration of the analyte. Assuming there are R analytes, the data held in the $I \times J$ matrix $\mathbf{X}$ can be modeled by R bilinear components as

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{b}_r \mathbf{a}_r^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where $\mathbf{b}_r$ is an I-vector holding the concentration of the rth analyte in the I samples, $\mathbf{a}_r$ is the time profile of the rth analyte, and the matrix $\mathbf{E}$ holds the residual variation. For each sample the time profile is described as a sum of the individual profiles weighted by the corresponding concentrations of the analyte, $b_{ir}\mathbf{a}_r$. This model implies that the time profiles do not change from sample to sample. If the analytes are completely separated, the individual profiles can immediately be extracted, in which case no additional mathematical modeling is required. If the time profiles overlap, this corresponds mathematically to the vectors $\mathbf{a}_r$, $r = 1, \ldots, R$, being non-orthogonal. Resolving or rather estimating the profiles of the pure analytes in such a case has received a lot of attention in chemometrics, starting with the work of Lawton and Sylvestre.[1] Owing to the fundamental rotational indeterminacy in bilinear modeling, it is not possible to estimate the pure profiles from the data without employing some sort of external knowledge in the decomposition in order to obtain a unique model. The word ‘external’ is to be taken lightly here, since the necessary knowledge may sometimes be obtained directly from the data. The main way of obtaining uniqueness is to identify selective variables (or samples), i.e. elution times where only one analyte is present (or absent). As described theoretically in Reference 2, this may lead to a unique or partially unique decomposition. The presence of selective variables forms the basis for most traditional resolution techniques in chemistry. Another approach is based on the use of constraints. One may estimate the parameters in the bilinear model under constraints such as non-negativity of concentration estimates or unimodality of elution profiles.[3] While constraints are useful for improving the estimates of model parameters, they do not lead to uniqueness in general. Rather, they help reduce the feasible set of solutions.

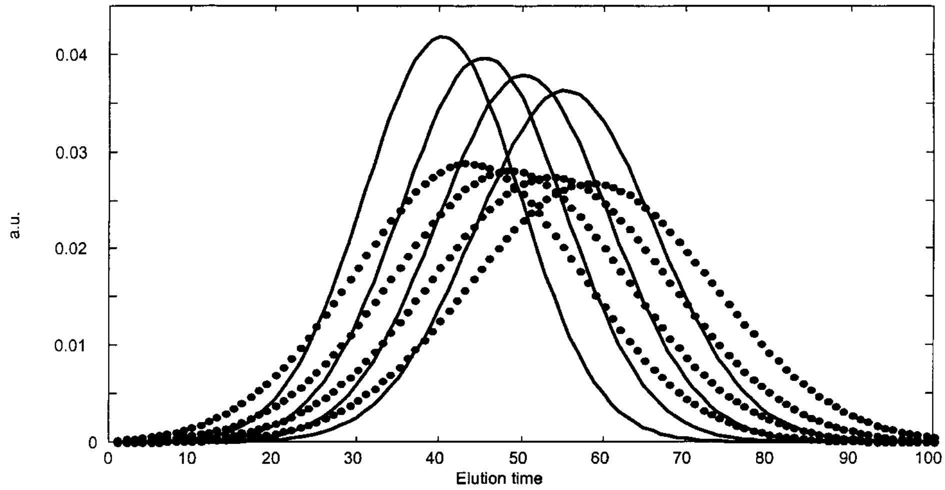

Figure 2. Elution profiles $\mathbf{F}_k$ from first experiment (full lines) and last experiment (dotted lines). The first and last experiments have the most dissimilar elution profiles, and the profiles change gradually throughout the experiments. Note that as the profiles shift, their width expands as well

When spectral detection rather than univariate detection is used, a three-way array is obtained, the third mode consisting of measurements at different wavelengths. It is well known that for *three-way* chromatographic data with no retention time shifts it is possible to resolve uniquely the underlying components without any additional constraints by the use of the PARAFAC1 model.[4] Thus the addition of a third spectral mode is highly convenient, since otherwise resolving the individual components may not be possible.

The primary concern in this paper is the problem of modeling three- and higher-way chromatographic data *with retention time shifts*. In the following we will first describe the chromatographic data and a set of simulated data used for introducing the PARAFAC2 model with respect to modeling retention time shifts. A short description of the possible models for resolving chromatographic multiway data is given. Finally the results of modeling the simulated as well as the real data are provided.

DATA

Simulated data

A three-way data set was generated for simulating spectrally detected chromatographic data with retention time shifts. Four analytes with overlapping chromatographic peaks were used. The data were generated according to the model

$$\mathbf{X}_k = \mathbf{F}_k \mathbf{D}_k \mathbf{A}^{\mathrm{T}} + \mathbf{E}_k \tag{2}$$

where $\mathbf{X}_k$ is the measured data from sample (i.e. experiment) k, $\mathbf{F}_k$ is a 100×4 matrix holding the elution profiles of the four (fictitious) analytes present in sample k, $\mathbf{D}_k$ is a 4×4 diagonal matrix holding the concentrations of the four analytes in sample k in its diagonal, and the matrix $\mathbf{A}$ is a 30×4 matrix holding the spectra of the four analytes, chosen as in Figure 1. The matrix $\mathbf{E}_k$ holds the added noise. Thus only the spectra in $\mathbf{A}$ are constant over the samples. The data set consists of data from ten samples. In different samples the concentrations of the analytes were chosen randomly (evenly distributed between zero and one) and the elution profiles were shifted differently as described below. Thus the data array is of size 100 (time) $\times$ 30 (spectrum) $\times$ 10 (sample).

Normally distributed heteroscedastic noise was added proportional to the size of the signal such that the variance of the noise was 5% of the variance of the systematic variation. Note that this is a relatively large amount of noise.

The following choice of structure in $\mathbf{F}_k$ (containing the elution profiles) was used. In any specific experiment all elution profiles had identical shifts. The amount of shift was gradually increased from zero in experiment 1 (Figure 2, full lines) to four time units in experiment 10 (Figure 2, dotted lines). With increasing shifts the width of the peak area was also increased accordingly, being proportional to the square root of the elution time.

If the data fit the premises of the PARAFAC2 model, the PARAFAC2 model gives unique parameters (up to trivial scaling and permutations). Since the 'true' parameters (pure spectra, concentrations and elution profiles) will provide a model that also gives the best fit, the PARAFAC2 parameters will thus be estimates of the true parameters. This is quite dissimilar from bilinear modeling where the rotational invariance of the solution makes it impossible to estimate the parameters unless auxiliary information is available. However, in this case it is known *a priori* that the data do not fit the PARAFAC2 model perfectly. For this to hold, the cross-product of the matrix holding the elution profiles, $\mathbf{F}_k$, should be constant over k as elaborated on in Part I.[5] Thus $\mathbf{F}_k{}^{\mathrm{T}}\mathbf{F}_k = \mathbf{G}$ for any k. That this is not the case in the above example is easily shown from

$$\mathbf{F}_1^{\mathrm{T}}\mathbf{F}_1 = \begin{bmatrix} 1{\cdot}00 & 0{\cdot}82 & 0{\cdot}93 & 0{\cdot}96 \\ 0{\cdot}82 & 1{\cdot}10 & 0{\cdot}60 & 1{\cdot}01 \\ 0{\cdot}93 & 0{\cdot}60 & 0{\cdot}96 & 0{\cdot}80 \\ 0{\cdot}96 & 1{\cdot}01 & 0{\cdot}80 & 1{\cdot}05 \end{bmatrix}$$

and

$$\mathbf{F}_{10}^{\mathrm{T}}\mathbf{F}_{10} = \begin{bmatrix} 1{\cdot}00 & 0{\cdot}90 & 0{\cdot}96 & 0{\cdot}98 \\ 0{\cdot}90 & 1{\cdot}05 & 0{\cdot}77 & 1{\cdot}00 \\ 0{\cdot}96 & 0{\cdot}77 & 0{\cdot}98 & 0{\cdot}89 \\ 0{\cdot}98 & 1{\cdot}00 & 0{\cdot}89 & 1{\cdot}02 \end{bmatrix}$$

The cross-products shown above have been normalized by scaling the first element to the value one for easier comparison. It is readily seen that these matrices are not identical and hence the requirements for the PARAFAC2 model to hold are not valid here. Thus the PARAFAC2 model will not fit the data perfectly, though still give unique estimates of parameters. The crucial aspect is to investigate if PARAFAC2 is still a reasonable model to use and if it can provide sensible estimates of the underlying parameters (spectra, profiles and concentrations). It is less constrained than a corresponding PARAFAC1 model, hence it is the main hypothesis in this paper that it can be expected to perform better than PARAFAC1. We aim to show that for reasonable deviations from perfect data, PARAFAC2 will still provide good estimates of the underlying parameters.

Chromatographic data

Fifteen samples of thick juice from different sugar factories were introduced into a Sephadex G25 low-pressure chromatographic system using a 0·02 M NH_4Cl/NH_3 buffer (pH 9·00) as carrier. In this way the high-molecular reaction products between reducing sugars and amino acids/phenols are separated from the low-molecular free amino acids and phenols. The high-molecular substances elute first, followed by the low-molecular species. Aromatic components are retained the longest time owing to a high affinity to the Sephadex material. The sample size was 300 μl and a flow of 0·4 ml min^{-1} was used. Twenty-eight discrete fractions of 1·2 ml were sampled and measured spectrofluorometrically on a Perkin Elmer LS50B spectrofluorometer.

The column was a 20 cm long glass cylinder with an inner radius of 10 mm packed with Sephadex G25 fine gel. The water used was doubly ion exchanged and millipore filtrated upon degassing. The excitation–emission matrices were collected using a standard 10 mm × 10 mm quartz cuvette, scanning at 1500 nm min^{-1} with 10 nm slit widths in both excitation and emission monochromators (250–440 nm excitation, 10 nm intervals; 250–560 nm emission, 4 nm intervals). For each sample, 28 excitation–emission matrices are measured, one for each fraction collected. Thus the size of the four-way data set is 28 (fraction) × 20 (excitation) × 78 (emission) × 15 (sample).

METHODS

A structural model of chromatographic data will first be developed for the ideal situation in which there are no retention time shifts. Subsequently it will be shown how to accommodate this model for handling retention time shifts. First only three-way data will be considered and afterwards it will be shown how to extend the results to four-way data as well as the mathematical consequences of such an extension. Then the results of applying the PARAFAC2 and competing models to the simulated three-way and real four-way chromatographic data are shown.

Consider data such as the above-mentioned where fluorescence spectroscopy is used for detection. When the emission wavelength is fixed, then at each elution time an excitation spectrum is measured. This corresponds conceptually to the normal situation in UV-vis detection chromatography. Let x_{ijk} be the emission intensity of the ith fraction (elution time) of the kth sample measured at the jth excitation wavelength. For a dilute solution in which no quenching occurs it holds that this intensity is the sum of intensity contributions from the individual fluorophoric entities in the sample plus some additional noise. Assume there are R independent fluorophores. For each fluorophore r the emission intensity is linearly dependent on the concentration c_{kr} in the kth sample. It is also linearly dependent on the ‘quantum yield’ at excitation wavelength j, a_{jr}. Finally it is linearly dependent on the relative amount of sample present in the ith fraction, f_{ir}. Thus the model of the data can be stated as

$$x_{ijk} = \sum_{r=1}^{R} f_{ir} a_{jr} c_{kr} + e_{ijk} \tag{3}$$

This model may also be stated in terms of matrices. Let $\mathbf{X}$ be the $I \times JK$ matrix holding the $I \times J \times K$ three-way array with typical elements x_{ijk}. The first J columns of $\mathbf{X}$ correspond to the $I \times J$ slab obtained from the three-way array by setting k equal to one. The $I \times R$ loading matrix $\mathbf{F}$ holds the parameters f_{ir}, and $\mathbf{A}$ ($J \times R$) and $\mathbf{C}$ ($K \times R$) are defined likewise. The columns of $\mathbf{F}$ will be the estimated elution profiles, the columns of $\mathbf{A}$ the estimated spectra, and the elements in $\mathbf{C}$ the estimated concentrations. Then it holds that the PARAFAC1 model can be stated as

$$\mathbf{X}_k = \mathbf{F}\mathbf{D}_k\mathbf{A}^{\mathrm{T}} + \mathbf{E}_k \tag{4}$$

where $\mathbf{X}_k$ is the kth frontal slab of the three-way array and $\mathbf{D}_k$ is a diagonal matrix holding the kth row of $\mathbf{C}$ in its diagonal.

From the theory of the PARAFAC1 model[4,6] it immediately follows that given the appropriateness of the model it is possible to resolve the data into meaningful components pertaining to individual analytes. This is so because the PARAFAC1 model is uniquely identified up to scaling and permutation of the components under mild conditions.[7–10] The model of the chromatographic data derived above assumes that the elution profiles of individual components, i.e. the columns $\mathbf{f}_r$ of $\mathbf{F}$, are identical in each sample. However, this is not the case in the presence of retention time shifts. In such situations, using the PARAFAC1 model will be problematic. We then have to replace the first mode loadings $\mathbf{F}$ with a set of loadings $\mathbf{F}_k$ specific to sample k. The elution profiles $\mathbf{F}_k$ for a specific experiment k are then unrelated to the profiles from another experiment, so as to allow for retention time shifts in the model. A model of shifted data may therefore generically be stated as

$$\mathbf{X}_k = \mathbf{F}_k\mathbf{D}_k\mathbf{A}^{\mathrm{T}} + \mathbf{E}_k \tag{5}$$

The parameters and residuals in this model are different in general from the ones given in equation (4), but the matrices are given the same names in order to stress that ideally these should be identical. This model is problematic for several reasons. First of all it possesses no uniqueness properties in the sought sense since it can be shown to be equivalent to a bilinear model of the data unfolded to a two-way matrix. Also important, though, is that it assumes no relation at all between equivalent elution profiles in different samples. If the elution profiles *are* somehow related, not using this will lead to an unnecessarily high uncertainty in the estimated components.

Between the two extremes of having all $\mathbf{F}_k$ equal to $\mathbf{F}$ and having $\mathbf{F}_k$ unconstrained there are several possibilities for imposing structure in $\mathbf{F}_k$. It is the choice of the structure of $\mathbf{F}_k$ that determines the structure of the model. The PARAFAC2 model offers one such intermediate model. In the PARAFAC2 model each loading matrix $\mathbf{F}_k$ is modeled as

$$\mathbf{F}_k = \mathbf{P}_k\mathbf{F}, \quad k = 1, \ldots, K \tag{6}$$

where $\mathbf{P}_k$ is an $I \times R$ column-wise orthonormal matrix and $\mathbf{F}$ is of size $R \times R$. The matrix $\mathbf{F}$ represents the common part of the elution profile matrices from different experiments in an R-dimensional subspace, while the matrix $\mathbf{P}_k$ determines the specific manifestation of these profiles in the I-dimensional space of the kth experiment.[†] One may of course also envision other ways of imposing structure in $\mathbf{F}_k$, but it seems that this type of structure is adequate for approximating many occurring deviations from the strict linearity required in the standard PARAFAC1 model. A very important feature of the PARAFAC2 model is that it retains the advantage of intrinsic structural uniqueness as discussed at length in References 5, 11 and 12.

The structure imposed in $\mathbf{F}_k$ can also be formulated differently by observing that equation (6) is equivalent[5] to requiring

$$\mathbf{F}_k^{\mathrm{T}}\mathbf{F}_k = \mathbf{F}^{\mathrm{T}}\mathbf{F}, \quad k = 1, \ldots, K \tag{7}$$

This means that for every sample k a set of elution profiles $\mathbf{F}_k$ is estimated under the constraint that the cross-products of the profile matrix are identical. It is simple to show that if for example the profiles of all analytes are shifted the same amount, if there is no peak broadening and the elution baseline is represented both before and after all analytes appear, then this assumption will be valid. If these assumptions are not met, the PARAFAC2 model is still less restrictive than the PARAFAC1 model while being unique. Thus even data that do not conform exactly to the restrictions may be better

[†] Note that the matrix $\mathbf{F}$ appearing in the PARAFAC2 model is not of the same size as the one appearing in the PARAFAC1 model.

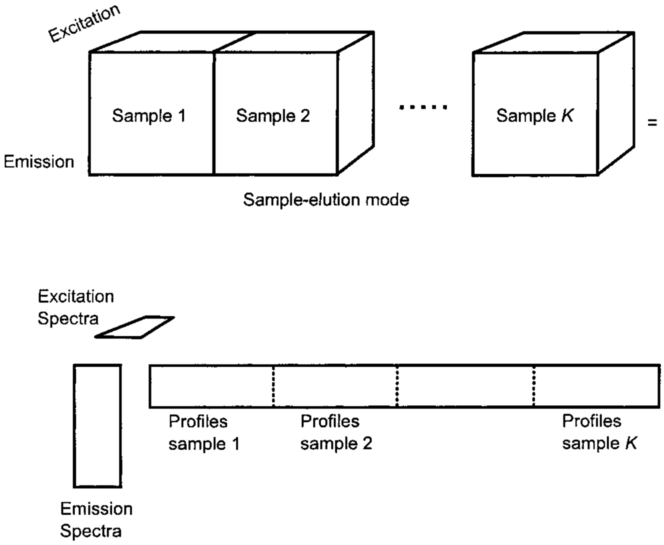

Figure 3. Four-way chromatographic data represented as a three-way array where sample and elution profile modes are combined into one. Below the corresponding three-way PARAFAC1 model is depicted, showing that for this unfolding the PARAFAC1 model estimates the elution profiles from each sample independently

modeled by PARAFAC2 than by PARAFAC1, since the model misspecification will be less pronounced for PARAFAC2.

Having discussed the three-way version of the PARAFAC2 model, it is appropriate to discuss aspects of modeling four-way data. As discussed in References 3 and 5, the PARAFAC2 model is easily extended to higher orders. An interesting aspect of the four-way model is that even if no constraints are imposed on $\mathbf{F}_k$, the model will still be unique, since the four-way model with unconstrained $\mathbf{F}_k$ is equivalent to a three-way PARAFAC1 model of the four-way data unfolded to a three-way array.[‡] Since the chromatographic data are four-way, it is therefore possible to validate the four-way PARAFAC1 and PARAFAC2 models against the results of the three-way PARAFAC1 model fitted to the unfolded four-way data. Regardless of the presence of retention time shifts the three-way PARAFAC1 model will give reasonable estimates of the model parameters if the elution and sample modes are combined in the unfolding (Figure 3).

Determining the model complexity

For PARAFAC1 as well as PARAFAC2 it is essential to use the correct number of components. In two-way analysis this is also important, but for multiway models the importance is even more pronounced. In most two-way analyses one is mainly interested in determining a suitable subspace, while in PARAFAC models the specific orientation within the subspace is also important. Moreover, PARAFAC models are not nested, so choosing e.g. a four-component model instead of a three-component model has implications not only for the additional component but also for the orientation of *all* four components.

[‡] Still, if the added structural constraint of the PARAFAC2 model is valid, it is preferable to use it, since added constraints (on $\mathbf{F}_k$) will in general provide more robust and precise parameter estimates.[3]

In order to determine the proper number of components for PARAFAC1 as well as PARAFAC2 models, several possibilities exist. As for ordinary two-way principal component analysis, methods based on judging residuals and on resampling are possible. For multiway models, however, some additional tools are available that are very helpful in determining the proper number of components. The split-half analysis[3] is founded on exploiting the uniqueness properties of the PARAFAC1 and PARAFAC2 models. If the right number of components is chosen, the 'true' underlying latent variables will be found. This will hold regardless of which samples are used for estimating these. If the proper number of components is not used, the estimated parameters will be linear combinations of the true parameters and therefore depend on which samples are used.

Another powerful tool for assessing the model complexity of PARAFAC1 models is the core consistency diagnostic suggested in Reference 3 and elaborated on in detail in Reference 13. It is based on the fact that the PARAFAC1 model can be posed as a restricted Tucker3 model where the core array is fixed to be a superidentity array.[14] The core consistency diagnostic amounts to first calculating the optimal unconstrained core array for a Tucker3 model where the loading matrices are the ones obtained by the PARAFAC1 model at hand. Then the core consistency diagnostic given as a percentage is defined as

$$\text{core consistency } = 100\left(1-\frac{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}(g_{def}-t_{def})^2}{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}t_{def}^2}\right) \tag{8}$$

where g_{def} and t_{def} denote the elements of the calculated core and of the intrinsic superdiagonal core respectively. If $\mathbf{G}$ is equal to $\mathbf{T}$, the core consistency is perfect and has a value of unity (100%), which indicates that the PARAFAC1 model at hand is indeed appropriate. At the other extreme the consistency may be below zero if the PARAFAC1 model is inappropriate or the variation is purely random, hence mostly off-superdiagonal.

As demonstrated in Reference 13, if the number of components in the hypothesized model exceeds the proper number of components, the Tucker3 core array will deviate considerably from superdiagonality. This will not be the case if the proper number of components is used. Thus the highest number of components that maintains a sufficiently superdiagonal Tucker3 core array will be the adequate number of components to use.

RESULTS

Simulated data

The results of fitting PARAFAC1 and PARAFAC2 models to the simulated data using the correct number of components (i.e. four) are shown in Figure 4. The PARAFAC2 estimates are closer to the true values than the PARAFAC1 estimates. Furthermore, it can be seen that the PARAFAC2 estimated elution profiles are less smooth than the corresponding PARAFAC1 estimates. This is an indirect illustration of the important property of PARAFAC2 that it puts fewer restrictions on the elution profiles. This is needed because such restrictions are infeasible when there are shifts. In this case, where a substantial amount of noise was added to the data, the estimated elution profiles become rather unsmooth, but they do follow the original profiles closely.

In order to verify that the PARAFAC2 model is superior to the PARAFAC1 model for the given data, 100 simulations were performed according to the above data but with different random concentration matrices. For every simulated data set the two models were fitted and the correlations

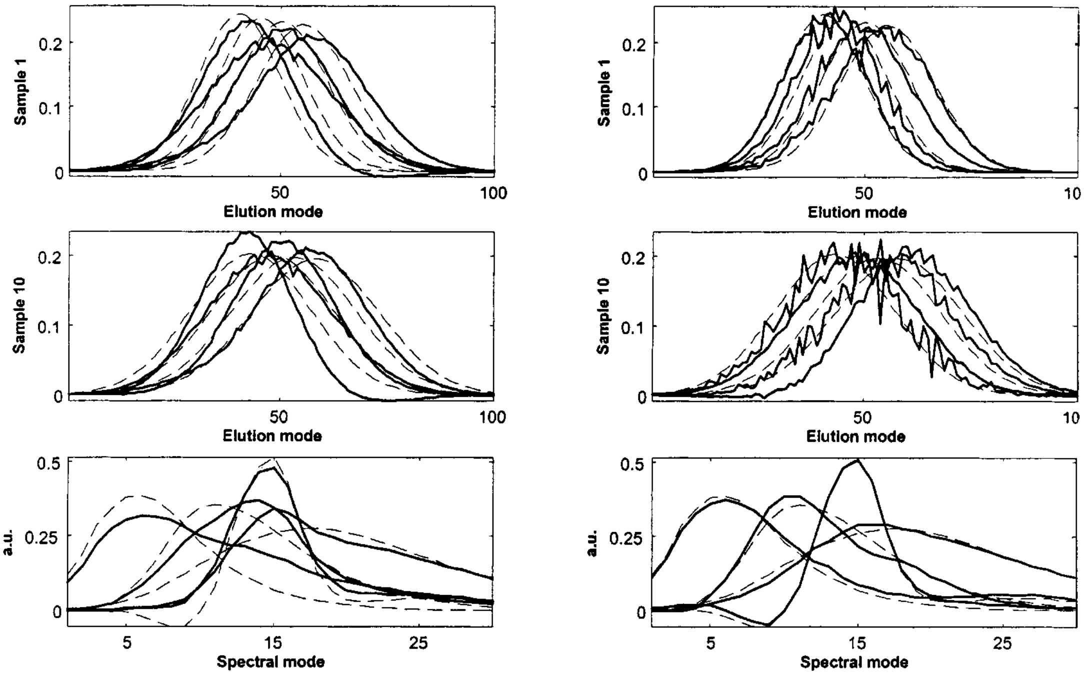

Figure 4. Estimated profiles and spectra from simulated data. The top plots show the true profiles (broken lines) together with estimates (full lines). PARAFAC1 estimates are to the left and PARAFAC2 estimates to the right. The middle plots show the same for experiment 10 and the bottom plots shows the reference spectra (broken lines) compared with the estimates

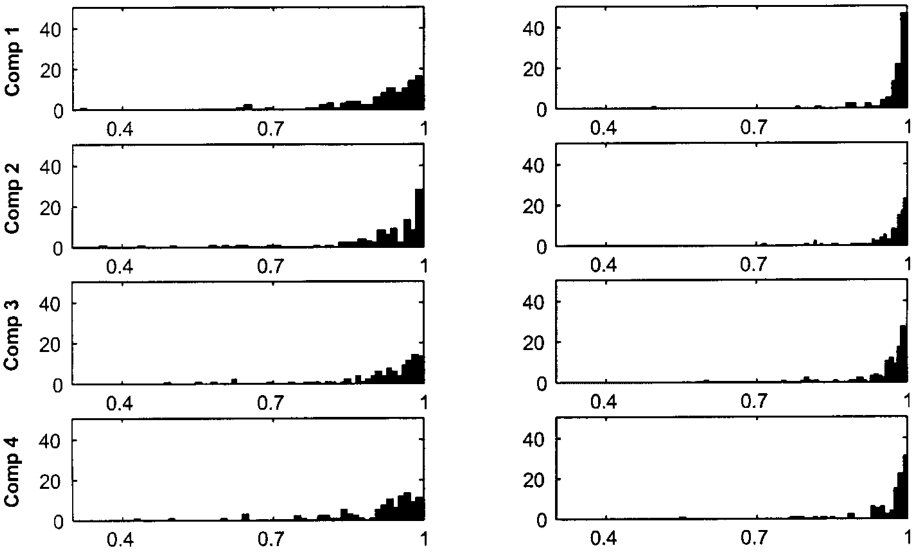

Figure 5. Histograms showing correlation between estimated and true concentrations for each component for PARAFAC1 (left) and PARAFAC2 (right). The histograms are based on 100 different models with different random concentration matrices

Table 1. Explained variation and core consistency for different three-way PARAFAC1 models, of chromatographic data

Number of components	Explained variation (%)	Core consistency (%)
3	96·4	91·9
4	98·9	96·3
5	99·3	20·6
6	99·4	15·1

between estimated and true concentrations calculated. In Figure 5 these correlations are shown. Each plot is a histogram containing the absolute correlation between the estimated and true concentrations of one specific analyte for one specific model over all 100 data sets. It is evident that the PARAFAC2 model is generally superior to the PARAFAC1 model. The correlations between true and estimated concentrations for the PARAFAC2 model are much more skewed towards one than for the PARAFAC1 model.

Chromatography

The first step in modeling the chromatographic data is to determine how many components to use in the model. In order to establish the correct number of components, a three-way PARAFAC1 model was investigated in which the sample and elution modes were concatenated into one mode (see Figure 3). In this way, retention time shifts will not affect the model, since the profiles of each sample will be modeled independently. For three-, four-, five- and six-component models the core consistency (equation (8)) as well as the percentage of variation explained was calculated. The percentage of explained variation was defined as

$$\text{variation explained} = 100\left(1-\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}(x_{ijkl}-m_{ijkl})^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}x_{ijkl}^2}\right) \tag{9}$$

where x_{ijkl} is an element of the four-way array and m_{ijkl} is the corresponding element of the model of the array.

For the posed models the results are given in Table 1. Note that based on the percentage of variation explained, it is difficult to assess which of the four candidate models is the most preferable since they all explain approximately the same amount of variation. Using the core consistency, however, the picture is much clearer. Three- and four-component models are seen to be suitable since they both have very high consistencies. A five- or six-component model is definitely not appropriate, since the loading matrices that should reflect the subspace of the systematic variation are mainly descriptive of variation on the off-superdiagonal part of the array (indicated by the low core consistency). Since four is the highest number of components for which the model assumptions hold, it may be concluded that four components provide an adequate model complexity of the given data under the premises of the PARAFAC1 model.

Having established the number of components to use, the two competing four-way models of the data were fitted: a four-way four-component PARAFAC1 model and a four-way four-component PARAFAC2 model. For both models, non-negativity was imposed on all parameters except the

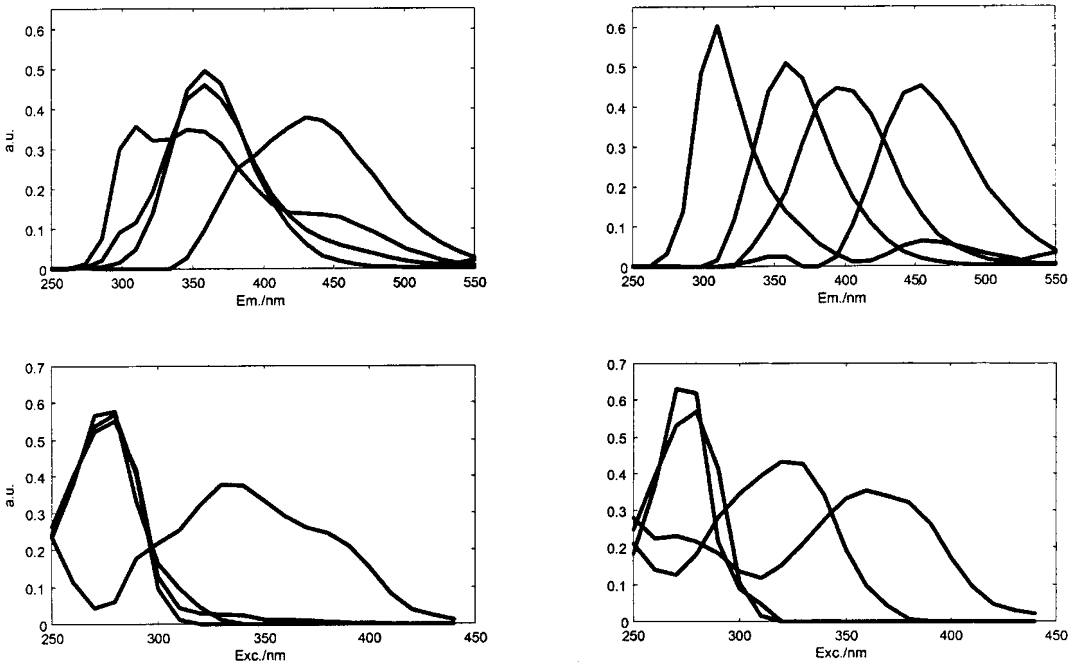

Figure 6. Estimated emission (top) and excitation (bottom) spectra from four-way PARAFAC1 (left) and four-way PARAFAC2 (right)

elution profiles in PARAFAC2, since imposing non-negativity on these is difficult.[5] In Figure 6 (left) the excitation and emission mode loadings of a four-component PARAFAC1 model are shown. The parameters are not very appealing. The alikeness of several components suggests that the model may not be valid. However, the solution is stable in the sense that it was obtained several times from different starting values. In Figure 6 (right) the excitation and emission mode loadings of a non-negativity-constrained PARAFAC2 model are also shown. These parameters look reasonable and are very different from the PARAFAC1 loadings, especially in the emission mode. The PARAFAC2 model seems to be better. Based on these results alone, it is difficult, though, to conclusively claim that the PARAFAC2 model is valid and better than the PARAFAC1 model.

A very simple way of validating which model is better admits itself as mentioned before. The sample and elution modes may be combined into one mode and the subsequent three-way array uniquely modeled by a three-way PARAFAC1 model. Since each elution mode will then be modeled separately for each sample, possible retention time shifts will not affect the appropriateness of the model.

For the model of the three-way data the excitation and emission mode loadings are shown in Figure 7. Note the close similarity between the three-way PARAFAC1 and four-way PARAFAC2 solution. All three models (three-way PARAFAC1, four-way PARAFAC1 and four-way PARAFAC2) should theoretically be identical if no retention time shifts are present. Since the four-way PARAFAC1 model gives substantially different parameter estimates, it may safely be concluded that this model does not fit the characteristics of the data. The most likely reason for this is retention time shifts.

From the three-way model of the data a set of loadings is also obtained in the combined elution/sample mode. Reshaping the loading for one specific component to a matrix, a set of elution profiles for this 'analyte' is obtained, one for each sample. In Figure 8 this is shown for component 1. These

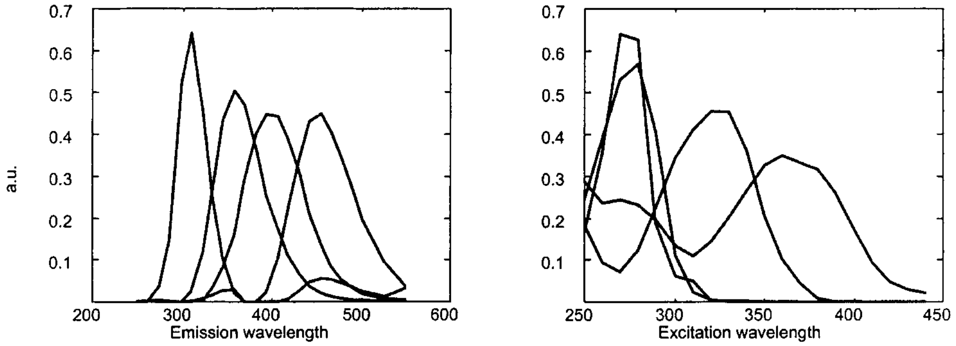

Figure 7. Emission (left) and excitation (right) mode loadings estimated from three-way non-negativity-constrained PARAFAC1 model

estimated profiles are not subjected to model error due to retention time shifts, since they stem from the three-way model.

It is readily seen that even though the elution profiles should be identical in each run, this is certainly not the case. There are huge shifts in the retention times from sample to sample, probably caused by the very different contents of the samples. This explains why four-way PARAFAC1 cannot fit these data well. The gel in the column is known to be sensitive toward the concentration of phenolic compounds and certain amino acids. The inter-sample variation in the elution profiles is probably due to different contents of such compounds with high affinity for the chosen gel causing the shifts in retention times.

It is interesting to compare the elution profiles estimated by three-way PARAFAC1 with the

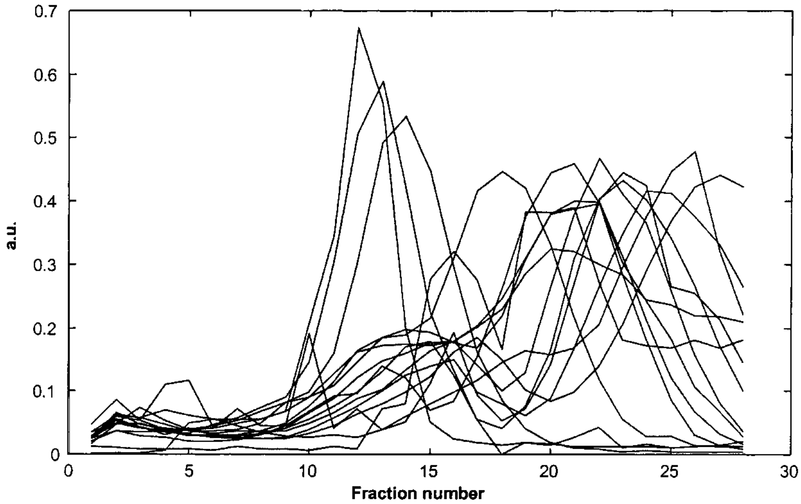

Figure 8. Estimated elution profiles of component 1 (not scaled) estimated from a three-way PARAFAC1 model. Each line is the estimated profile of the component in one specific sample. If no retention time shifts were present, all profiles should be identical!

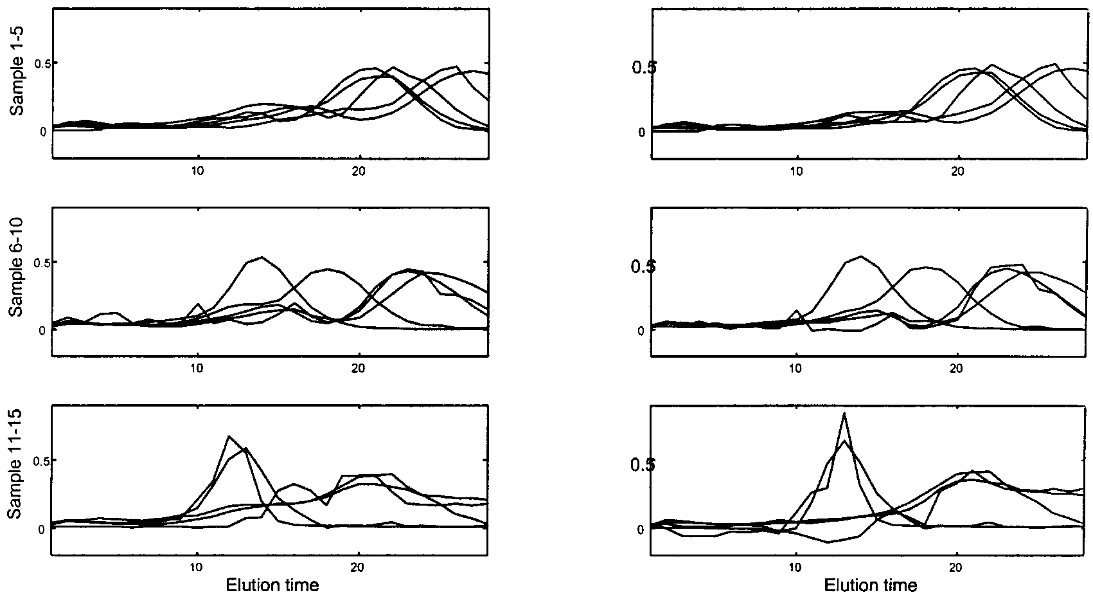

Figure 9. Estimates of elution profiles of component 1 in 15 different samples. Estimates from three-way PARAFAC1 are shown to the left and from four-way PARAFAC2 to the right. The top plots show the estimates of the first five samples, etc.

estimates obtained from PARAFAC2. As for the three-way model and unlike the four-way PARAFAC1 model, PARAFAC2 provides individual profiles for each sample ($\mathbf{F}_k$). In Figure 9 the estimated profiles of component 1 in all samples are compared for three-way PARAFAC1 and four-way PARAFAC2. Note that the PARAFAC1 elution profiles are identical to the ones shown in Figure 8. As for the spectral parameters the similarity is very high even though the deviations between the 15 elution profiles are not of a type expected to be perfectly modeled by PARAFAC2.

Performing a split-half analysis for both the four-way PARAFAC1 and the PARAFAC2 model substantiated that the four-way PARAFAC1 model is not suitable, since the parameters did not replicate over different subsets. The data were divided into two groups by assigning eight samples to one group and seven to another. For both subsets a PARAFAC1 and a PARAFAC2 model were fitted. In Figure 10 the resulting emission and excitation mode loadings are shown. There are large discrepancies in the PARAFAC1 parameters depending on which subset is used, while for the PARAFAC2 model these discrepancies are smaller and probably caused by the very low sample size (seven and eight respectively).

CONCLUSION

In this application a suggestion has been given for the solution of a very important and frequently arising problem, namely shifted data. It has been shown that even though the data are severely shifted, PARAFAC2 apparently is capable of modeling the data. In this case, validation could be very elegantly performed by unfolding the four-way data to a three-way structure for which the PARAFAC1 model, and its ensuing uniqueness, holds. However, usually, shifted chromatographic data are at most three-way and therefore such a rearrangement in order to attain uniqueness is impossible. Furthermore, using the four-way PARAFAC2 model, more structure is imposed in the model than with the three-way PARAFAC1 model for the unfolded data, which is preferable from an

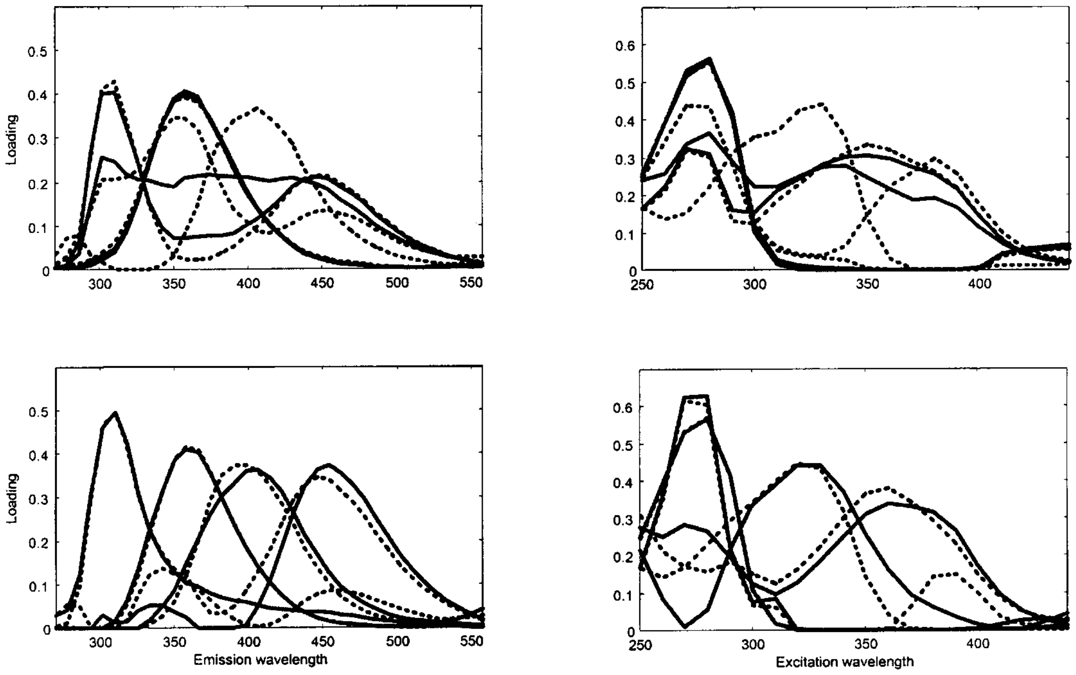

Figure 10. Split-half analysis. The top plots give the results from four-way PARAFAC1 and the bottom plots the results from PARAFAC2. The left plots show the emission mode parameters and the right plots the excitation mode parameters. Loading vectors estimated from a subset of eight samples are shown with full lines, and loading vectors estimated from a subset of seven samples are shown with dotted lines

interpretation as well as a noise reduction point of view.

The three-way PARAFAC2 model appears to provide a good approach for solving variable shifts for three-way data, and further applications to chromatographic data will help substantiate this conclusion.

ACKNOWLEDGEMENTS

For the first two authors, financial support was obtained from Professor Lars Munck, Food Technology, Department of Dairy and Food Science, Royal Veterinary and Agricultural University, Denmark through Nordic Industry Foundation project P93149 and FØTEK for part of the work reported here.

REFERENCES

1. W. H. Lawton and E. A. Sylvestre, *Technometrics.* **13**, 617 (1971).
2. R. Manne, *Chemometrics Intell. Lab. Syst.* **27**, 89 (1995).
3. R. Bro, *Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications,* University of Amsterdam, Amsterdam and Royal Veterinary and Agricultural University, Amsterdam (1998).
4. R. A. Harshman, *UCLA Working Papers Phonet.* **16**, 1 (1970).
5. H. A. L. Kiers, J. M. F. ten Berge and R. Bro, *J. Chemometrics.* **13**, 275 (1999).
6. R. Bro, *Chemometrics Intell. Lab. Syst.* **38**, 149 (1997).
7. R. A. Harshman, *UCLA Working Papers Phonet.* **22**, 111 (1972).
8. J. B. Kruskal, *Psychometrika.* **41**, 281 (1976).

9. J. B. Kruskal, in *Multiway Data Analysis*. ed. by R. Coppi and S. Bolasco, p. 8, Elsevier, Amsterdam (1989).
10. J. B. Kruskal, *Linear Algebra Appl.* **18**, 95 (1977).
11. J. M. F. ten Berge and H. A. L. Kiers, *Psychometrika.* **61**, 123 (1996).
12. R. Harshman and M. E. Lundy, *Psychometrika.* **61**, 133 (1996).
13. R. Bro and H. A. L. Kiers, '*A new efficient method for determining the number of components in PARAFAC models*', submitted (1999).
14. L. R. Tucker, *Psychometrika*, **31**, 279 (1966).